OPTIMISTIC MODEL SELECTION FOR EXPLORATION CONTROL

Wyatt, Jeremy[*]

***Abstract*--A large number of heuristics have been suggested for exploration control in reinforcement learning. We present a new model-based interval estimation algorithm that combines insights from a number of recent methods. Essentially it picks an optimistic model from the density over possible models and uses this to estimate the exploration value of each action.We show empirically that it outperforms current model-based methods on a maze task with respect both to performance while learning and to the quality of the policies obtained by the end of the learning period.**

***Index terms*—reinforcement learning, optimal control**

I. Introduction

The problem of how to act while learning is a class of optimal control problems with a long history [1][4][7]. In reinforcement learning (RL) it has taken the form of problems of (i) how to act so as to maximise performance during the learning agent's lifetime [9][11]; and (ii) how to act so as to identify a near optimal policy rapidly, or while accumulating acceptable costs [3][5][6][10]. These problems, while related, are not the same [16].

In Markov Decision Processes (MDPs) the optimal Bayesian solution to problem (i) is well known, but intractable [1][11]. Numerous approximations have been proposed. In this paper a new heuristic method is presented which brings together ideas from several recent approaches [9][10][15]. The algorithm derived is shown to outperform existing model-based techniques with respect to both problems (i) and (ii). The domain is a finite state MDP with an unknown transition function and a known reward function. All the methods considered in this paper are model-based. The paper is structured as follows. In Section II we describe the optimal solution as specified within a Bayesian framework, as originally outlined by [1][11], and recent heuristic approaches based on interval estimation (IE) . In Section III we describe the new algorithm which extends Wiering's [15] model based interval estimation (MBIE) method using a Dirichlet density. In Section IV we present the results of an empirical study, and show that the new algorithm outperforms both Wiering's and Meuleau's [12] exploration methods on a stochastic maze task.

II. Previous Work

We can distinguish four components of an exploration method: the measure of *local* exploratory value employed (e.g. reward based, counter based, error based, recency based, variance based); whether this measure is converted into a *distal* measure of exploratory value using a Bellman equation; whether the method for inferring the exploration value function is *model-based* or *model-free*; and what form the *decision rule* based on the exploration value function takes (eg ε-greedy, Boltzmann). All the methods considered here are model-based, and greedy (ie they always choose the best action according to the exploration value function). We first outline the optimal Bayesian specification of and solution to problem (i), for the case of an unknown MDP. All other exploration measures for problem (i) can essentially be considered an approximation to this.

The Bayesian approach is based on there being a space P of possible transition functions for the MDP , and a well-defined prior probability density over that space. The probability density over the space of possible finite MDPs for a known state space of S is constructed as follows. If state $i \in S$ has N possible succeeding states when action for a is taken, then the transition function from that state action pair is a multinomial distribution over the outcomes:

$$\vec{p}_i^{\,a} = \left\{ p_{i1}^a, p_{i2}^a \cdots p_{iN}^a \right\}$$

Eq. 1

The possible transition functions from i, a are the possible $\vec{p}_i^{\,a}$. We want a density over this space which is closed under sampling from any such multinomial. Martin [11] showed that the Dirichlet density has this property:

$$f\left(\vec{p}_i^{\,a} \mid \vec{m}_i^{\,a}\right) = \frac{\Gamma\left(\sum_{j=1}^{N} m_{ij}^a\right)}{\prod_{j=1}^{N} \Gamma\left(m_{ij}^a\right)} \prod_{j=1}^{N} \left(p_{ij}^a\right)^{m_{ij}^a - 1}$$

Eq. 2

[*] School of Computer Science, University of Birmingham, Birmingham, UK. B15 2TT. jlw@cs.bham.ac.uk

the density is parameterised by the $m_{ij}^{a} > 0$ for all j. The parameter vector is updated as follows, if a single observation of a transition $i \rightarrow_{a} j$ is made, then the new density is also Dirichlet with $m_{ij}^{a''} = m_{ij}^{a'} + 1$. The densities over the one step transition functions for all other state action pairs are independent. The density $f(\mathbf{P} \mid \mathbf{M})$ for a possible transition function $\mathbf{P} \in P$ for an MDP is therefore simply the product of the $f(\vec{p}_{i}^{a})$ over all i and *a*. This density is parameterised by the matrix $\mathbf{M} = [m_{ij}^{a}]$, where of. The additional information from a sequence of observations is captured in a count matrix $\mathbf{F}$. The posterior density given these observations is therefore simply parameterised by $\mathbf{M}'' = \mathbf{M}' + \mathbf{F}$. For convenience the transformation on $\mathbf{M}$ due to a single observed transition $i \rightarrow_{a} j$ is denoted $T_{ij}^{a}(\mathbf{M})$.

The value function in a Markov chain with unknown transition probabilities is thus itself a random variable, $\tilde{V}_{i}$. Given the usual squared error loss function the Bayesian estimator of expected return under the optimal policy is the expectation of $\tilde{V}_{i}$

$$V_{i}(\mathbf{M}) = \mathrm{E}[\tilde{V}_{i} \mid \mathbf{M}] = \int_{P} V_{i}(\mathbf{P}) f(\mathbf{P} \mid \mathbf{M}) d\mathbf{P}$$

Eq. 3

where $V_{i}(\mathbf{P})$ is the value of i given the transition function $\mathbf{P}$. When evaluated this is transformed into a known MDP defined on the information space $M \times S$:

$$\begin{aligned} V_{i}(\mathbf{M}) &= \max_{\mathrm{a}} \left\{ Q_{i}^{a}(\mathbf{M}) \right\} \\ &= \sum_{j} \bar{p}_{ij}^{a}(\mathbf{M}) \left(R_{ij}^{a} + \gamma V_{i}\left(T_{ij}^{a}(\mathbf{M}) \right) \right) \end{aligned}$$

Eq. 4

Where $\bar{p}_{ij}^{a}(\mathbf{M})$ is the marginal expectation of the Dirichlet. This shows how the Bayesian estimate of value elegantly incorporates the value of future information. The optimal solution to the well-known exploration-exploitation trade-off (problem (i) above) is thus to act greedily with respect to the Bayes Q-values. Because the solution involves dynamic programming over a tree of information states the problem is intractable. The simplest approximation to this is the certainty equivalent (CE) estimate, constructed by replacing $T_{ij}^{a}(\mathbf{M})$ with $\mathbf{M}$ in **Eq. 4**.

Approximate approaches to the exploration-exploitation trade-off typically circumvent this problem by some instantiation of the heuristic ‘‘be optimistic in the face of uncertainty” [2][9][12][13][14][15]. Most of these schemes calculate the uncertainty in some of the estimated quantities and add an exploration bonus based on this to a CE estimate of V_{i}. The first of these was Kaelbling’s interval estimation method [9][1]. This when applied to bandit tasks uses the upper bound of an interval estimate for the immediate reward associated with each action. The action selected is that with the highest upper bound on the immediate reward. When applied directly to Q-values in multi-stage decision problems this method uses a window or a decaying trace of previous Q-values to generate the estimate of the upper bound on the Q-values. This, however, means that the estimate picks up the non-stationarity in the Q-values due to their initial bias.

In addition the local exploration bonus is only combined with the estimated Q-values for action selection purposes. The bonus is not propagated to predecessor states and thus the resulting measure is local rather than distal. Meuleau and Bourgine [12] created a distal IE measure by combining the local IE bonus $\delta_{i}^{a}(\mathbf{M})$ with the reward so that it is propagated to predecessor states in the estimated model:

$$\xi_{i}^{a}(M) = \delta_{i}^{a}(\mathbf{M})(1-\gamma) + \sum_{j} \bar{p}_{ij}^{a}(\mathbf{M}) \left(R_{ij}^{a} + \gamma \max_{b} \left\{ \xi_{j}^{b}(\mathbf{M}) \right\} \right)$$

Eq. 5

where $\delta_{i}^{a}(\mathbf{M})$ is the local bonus, $(1-\gamma)$ is a scaling factor, and ξ_{i}^{a} is the exploratory value of taking action a in state i. The agent then follows a policy which is greedy with respect to ξ_{i}^{a}. One version of this algorithm (variance-based) also uses a window of previous Q values to calculate the local exploration bonus; while their worst case method uses an upper bound on the underlying variance in the return. Some form of asynchronous real time dynamic programming (ARTDP) is used to adjust the exploration value function on-line. Meuleau’s methods have been shown to outperform most current exploration techniques on a variety of tasks.

The approach taken by Wiering and Schmidhuber [15] is to extend the interval estimation concept in a different way. Rather than estimating the variance in the Q-values directly and using this to supply an exploration bonus, we can apply the optimism heuristic to the transition function $\mathbf{P}$. For each state action pair the upper bound of the $(1-\alpha)100\%$ confidence interval is calculated for the transition probability leading to the successor state with the highest estimated value. The other transition probabilities are renormalised, and ARTDP is applied to the optimistic MDP generated.

There are some minor drawbacks to this method. The algorithm uses a Gaussian density to model the uncertainty about each transition probability. In consequence the sample sizes for each transition have to be large before that assumption is justified. Because of

[1] Kaelbling’s inference method was model-free but Meuleau has introduced a model-based version. Given a model-based method it may be possible to define an estimate of expected error in the Q-value estimates that is not subject to such a bias. Such work would prove an interesting extension to Meuleau’s IEDP algorithms

this Wiering and Schmidhuber initially employ a distal counter-based exploration method to acquire a good estimated model, and then use their model-based interval estimation technique (MBIE) to bias the exploration to the most useful (highly rewarding) parts of the state space. The algorithm switches to MBIE when the changes over time in the value function become small. The result is a method that outperforms plain distal counter-based exploration both in terms of reward generated during the learning period; and in terms of the quality of the policy learned.

III. Improving Model-Based Interval Estimation

Since the Dirichlet is the natural conjugate density for sampling from a multinomial distribution, it seems appealing to employ this, rather than a Gaussian in calculating the interval estimates for the one step transition probabilities. This allows us to correctly and simply incorporate prior knowledge about the transition function. The main problem is how to choose the prior parameter matrix $\mathbf{M}$ in the case where we have no prior knowledge about $\mathbf{P}$. Since we do not know which transitions are possible we have two choices. Either we can assign a uniform (non-zero) prior to all possible transitions; or we can assign a prior to some subset. We take the second case to the limit by using a single additional terminal state k to represent possible unobserved transitions. A similar idea has been employed by Kearns and Singh [10]. In addition by making this state highly rewarding [13] we can induce a distal exploration value function that will drive the learner toward novel state action pairs.

The parameter matrix $\mathbf{M}$ is thus initially all zero except for a single hypothesised transition to the terminal state from every state action pair i, a the prior for which is m^a_{ik}. Initially $V_i = R_i$ for all states other than *k*. V_k is set to be some upperbound on the value function. Each time t the agent selects an action a in state i that maximises ξ^a_i and observes the transition $i \rightarrow_a j$. It then updates the parameter matrix $\mathbf{M}$ in the standard way. Because of our degenerate prior, each time a novel transition (in terms of the prior) is observed this update is not Bayesian since observations are incorporated for previously excluded hypotheses. Subsequent updates for an observed transition follow Bayes rule. Given the new information we then re-calculate the upper bound of the $(1-\alpha)100\%$ confidence interval for the transition probability to the state k for each action. The marginal density required for this computation is simply a Beta density, always following $\mathrm{Beta}\left(\mathrm{m}^{\mathrm{a}}_{\mathrm{ik}}, \sum_{j\neq k} \mathrm{m}^{\mathrm{a}}_{\mathrm{ij}}\right)$. Since we know the single prior parameter m_k and also that all other parameters are integer we can calculate the upper bounds for a suitable parameter set off-line. The other probabilities are then renormalised. The result is an optimistic MDP $\mathbf{P}_{\mathbf{opt}}$. An ARTDP method can then be applied to $\mathbf{P}_{\mathbf{opt}}$ to revise our estimate of the value function. The relevant Bellman equation is:

$$\xi^a_i(\mathbf{M}) = \sum_j \hat{p}^a_{opt,ij}(\mathbf{M})\left(R^a_{ij} + \gamma \max_b \left\{\xi^b_j(\mathbf{M})\right\}\right)$$

Eq. 6

where $\hat{p}^a_{opt,ij}(\mathbf{M})$ are the transition probabilities according to $\mathbf{P}_{\mathbf{opt}}$.

What behaviour should we expect the algorithm to exhibit? Initially exploration will be random as ties are broken randomly between novel actions in the current state. Once all actions have been tried in a state the measure will be a mix of a distal counter-based measure, and the maximum likelihood value function. Exploration will initially be weighted in favour of the former and move to the latter as the number of trials of each action in a state rises. This means that a high value for m_k will cause highly exploratory behaviour for a long period of time, whereas a low value will mean a rapid shift to exploitation.

This method can be seen as a relation of Kearns and Singh's E^3 algorithm in which the learner chooses either to identify the model by taking actions that drive it toward the unknown state set, or to exploit within the set of known states. In our algorithm as soon as a state action pair is tried it is considered known, and can thus be used in exploitation if it is appealing enough. Alternatively the algorithm can be seen as an extension of Wiering and Schmidhuber's method which uses a more appealing density to represent uncertainty in the model; and utilises this density in exploration control from the outset, providing a smooth -- rather than a sudden -- switch from distal counter based to MBIE exploration.

IV. Empirical Study

The task chosen to compare algorithms was a 400 state MDP maze task, similar to that in [15], shown in Figure. The starting state is in the centre of the maze (x,y = 11,10). There are four actions (N,S,E,W) and transitions have a likelihood of 0.8 of succeeding, 0.08 of carrying the agent laterally to the intended direction of travel, and 0.04 of carrying the agent one step in the opposite direction. Reward is a deterministic function of state. The four corners are terminal states, the top right corner generating a reward of 100, and the other three rewards of 50 each. The maze is filled with penalty fields (-4 or -1) and walls. The transition probabilities for actions that lead to walls are redirected into the state in which the action was taken, but these carry no penalties. was set in all experiments to be 0.95.

Three algorithms were tested: Wiering's model based interval estimation (Wiering's MBIE); Meuleau's variance based IEDP+ [12]; and our Dirichlet based MBIE. Each algorithm was optimised across a range of parameters. Each run of an algorithm consisted of approximately

25000 time steps, and possibly of many trials. When the agent reaches a terminal state a new trial is started by sending the agent back to the starting state. The last trial in each run is allowed to terminate even if it means the total run length exceeds 25000 steps. Each algorithm was tested for 100 runs. The main algorithm parameters were set as shown in Table 1. All algorithms were run using Wiering and Schmidhuber's version of prioritized sweeping [15], with the threshold for the priority queue, $\varepsilon = 0.001$, and the maximum number of backups per step $U_{\max} = 20$.

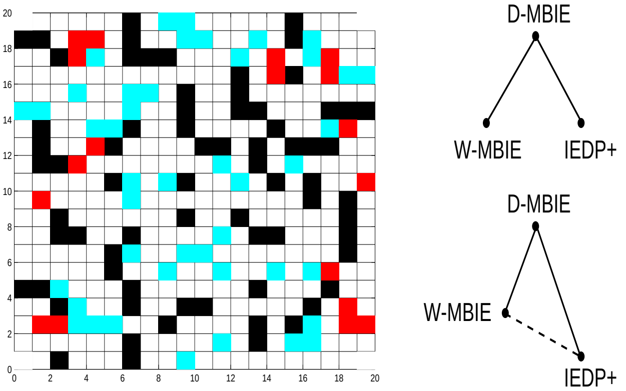

Figure 1 (a) A stochastic maze. Walls are marked in black, and penalty fields of -4 and -1 in dark and light grey respectively. (b) Partial order dominance for (top) expected regret and (bottom) reward generated. A solid line indicates significance at the 0.1% level, a dashed line indicates significance at the 1% level.

Algorithm	Parameter settings
IEDP+	Window length = 30 $\delta_1 = 20,60,100,400,1000$ $= \sigma_{\max}\left(2\delta_0(2) - \delta_0(3)\right)$
Wiering's MBIE	switching parameter $\eta = 2^{\beta}$ $\beta = 4,3,2,0,-1,-2,-3,-4,-5$ $K = 50, \alpha = 0.05$
Dirichlet MBIE	$m_{ik}^{a} = 2^{\beta}$ $\beta = 0,-1,-2,-3,-4,-5,-6,-7,$ $\alpha = 0.05$ $V_k = 100$

Table 1: Algorithm parameters

We measured the performance of each algorithm according to two performance criteria. The first is the total reward generated over the length of a run, averaged over all 100 runs. This is used following Meuleau [12] because it provides a finer discrimination between algorithms than the cumulative discounted reward. The expected value of the latter measure is that optimised by the Bayesian solution, so our performance criterion for exploration control of type (i) is different to the classic measure. The second performance measure assesses the expected regret of a greedy policy generated from the final agent model. In order to find the policy for each agent we extract the maximum likelihood transition model, and apply synchronous value iteration. The set of greedy actions is extracted for each state, and a greedy policy is created that breaks ties between actions randomly. The expected regret of this policy is calculated using the known MDP. Finally we present the regret in the starting state as a proportion of the value of an optimal policy in that state[2]. This measures performance of exploration control of type (ii). We present the results for the optimised parameters only for each algorithm on each criterion in Table 2, giving the average $\bar{x}$ and the sample standard deviation s for each measure. Dirichlet based MBIE outperforms both other methods for the optimal parameter settings on both performance criteria. The differences shown are significant at the 0.1% or 1% levels, using a 2 tailed t-test, with the partial dominance ordering shown in Figure 1(b).

Alg'ms	Best Param' values	Total reward		Expected reward	
		$\bar{x}$	s	$\bar{x}$	s
IEDP+	$\delta_1 = 60$	51082	2623	0.047	0.054
	$\delta_1 =1706$	-4710	445	0.034	0.039
Wiering's MBIE	$\beta =-1$	52684	3298	0.028	0.07
	$\beta =-5$	11311	654	0.022	0.023
Dirichlet MBIE	$\beta =-6$	56932	3478	0.078	0.064
	$\beta =-4$	22708	931	0.01	0.011

Table 2: Results

V. Discussion

We have presented a new heuristic algorithm which on a typical task outperforms two of the leading model-based explorers, which have themselves been shown to outperform almost all other widely used exploration techniques. The performance of all the algorithms varies significantly across the parameter set. A key question is therefore whether limited problem knowledge, e.g. the likely density of connections, can be used to guide parameter selection. We are testing the current algorithm against others on a range of standard tasks and randomly generated MDPs to try to establish this. There are a number of extensions to be made to the algorithm. First a different method could be used to calculate $\mathbf{P}_{\text{opt}}$ [8], this scheme would have the consequence that the agent could be optimistic about transitions to several high value neigbouring states. Our next goal, however, is to use the method for the construction of optimistic multi-time models in hierarchical reinforcement learners in order to guide exploration choices between options in Semi-MDPs. Finally we hope to generalise the technique to stochastic process models with function approximation.

[2] We use the a value of 0 as the baseline.

VI. BIBLIOGRAPHY

[1] R.E. Bellman. “Adaptive Control Processes: A Guided Tour”, Princeton University Press, 1961.

[2] P. Dayan and T. Sejnowski, “Exploration bonuses and dual control”, Machine Learning, 25, pp.5—22, 1996.

[3] R. Dearden, N. Friedman, and D. Andre, “Model-based Bayesian exploration”, Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI—99), Morgan Kaufmann, San Francisco, CA, pp.150—159,1999.

[4] A. A. Fel'dbaum, “Optimal Control Systems”, Academic Press, New York, 1965.

[5] C.N. Fiechter, “Efficient reinforcement learning”, Proceedings of the 7th Annual ACM Conference on Computational Learning Theory, ACM, pp. 88—97, 1994.

[6] C.N. Fiechter, “Expected mistake bound model for on-line reinforcement learning”, Proceedings of the 14th International Conference on Machine Learning, D.H. Fisher, editor, Morgan Kaufmann, pp. 116—124, 1997.

[7] J.C. Gittins, “Multi-armed Bandit Allocation Indices”, John Wiley & Sons, 1989.

[8] R. Givan, S. Leach, and T. Dean, “Bounded parameter Markov decision processes”, Recent Advances in AI Planning: 4th European Conference on Planning S. Steel and R. Alami, editors, Springer Verlag, 1997.

[9] L. Kaelbling, “Learning in Embedded Systems”, Ph.D. thesis, Dept of Computer Science, Stanford, 1990.

[10] M.Kearns and S.Singh, “Near-optimal reinforcement learning in polynomial time”, Proceedings of the Fifteenth International Conference on Machine Learning, J. Shavlik, editor, Morgan Kaufmann, pp. 260—268, 1991.

[11] J.J. Martin, “Bayesian Decision Problems and Markov Chains”, Wiley, New York, 1967.

[12] N.Meuleau and P.Bourgine, “Exploration of multi-state environments: Local measures and back-propagation of uncertainty”, Machine Learning, 35, pp.117—154, 1999.

[13] A.W. Moore and C.G Atkeson, “Prioritised sweeping: Reinforcement learning with less data and less time”, Machine Learning”, 13, pp.103—130, 1993.

[14] R.S. Sutton, Integrated architectures for learning, planning, and reacting based on approximating dynamic programming, Machine Learning: Proceedings of the Seventh International Conference on Machine Learning, B.W. Porter and R.J. Mooney, editors, pp. 216—224. Morgan Kaufmann, 1990.

[15] M.Wiering and J.Schmidhuber, “Efficient model-based exploration”, From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior, R.Pfeiffer, B.Blumberg, J.Meyer, and S.W. Wilson, editors, 1998.

[16] J.L. Wyatt, Exploration and Inference in Learning from Reinforcement, Ph.D. thesis, Dept. of Artificial Intelligence, University of Edinburgh, 1997.